Acta Informatica (2011) 48:51–66
DOI 10.1007/s00236-010-0131-6

ORIGINAL ARTICLE

Efficient systematic clustering method for k-anonymization

Md. Enamul Kabir · Hua Wang · Elisa Bertino

Received: 27 July 2009 / Accepted: 7 December 2010 / Published online: 12 January 2011

Abstract This paper presents a clustering (Clustering partitions record into clusters such that records within a cluster are similar to each other, while records in different clusters are most distinct from one another.) based k-anonymization technique to minimize the information loss while at the same time assuring data quality. Privacy preservation of individuals has drawn considerable interests in data mining research. The k-anonymity model proposed by Samarati and Sweeney is a practical approach for data privacy preservation and has been studied extensively for the last few years. Anonymization methods via generalization or suppression are able to protect private information, but lose valued information. The challenge is how to minimize the information loss during the anonymization process. We refer to the challenge as a *systematic clustering problem* for k-anonymization which is analysed in this paper. The proposed technique adopts group-similar data together and then anonymizes each group individually. The structure of systematic clustering problem is defined and investigated through paradigm and properties. An algorithm of the proposed problem is developed and shown that the time complexity is in $O(\frac{n^2}{k})$, where n is the total number of records containing individuals concerning their privacy. Experimental results show that our method attains a reasonable dominance with respect to both information loss and execution time. Finally the algorithm illustrates the usability for incremental datasets.

Md. E. Kabir · H. Wang
Department of Mathematics and Computing, University of Southern Queensland,
Toowoomba, QLD 4350, Australia
e-mail: kabir@usq.edu.au

H. Wang
e-mail: wang@usq.edu.au

E. Bertino (✉)
Department of Computer Science and CERIAS, Purdue University,
West Lafayette, IN, USA
e-mail: bertino@cs.purdue.edu

1 Introduction

In recent years, the phenomenal advances in technological developments in information technology have lead to an increase in the capability to store and record personal data about customers and individuals [2]. Data mining is a common methodology to retrieve and discover useful hidden knowledge and information from personal data. This has lead to concerns that personal data may be breached and misused. Therefore it is necessary to protect personal data through some privacy preserving techniques before conducting data mining.

One of the most important concepts for privacy is anonymity. Anonymity refers to a state where one's identity is completely hidden, and anonymity is oftentimes used as a synonym for privacy [3]. Anonymous data can protect individuals in two ways: firstly to protect identity privacy for example by making it impossible to learn to whom a data record is related and secondly, through attribute privacy for example making it impossible to know about a particular property of individuals. In any database, specially where health records are collected by hospitals or government organizations, anonymity has a significant role to protect privacy as the information linked to individuals could be highly sensitive. In commercial databases where organizations would like to disclose an individual's data to third parties (e.g. external organizations), anonymity could be used to protect the privacy of individuals as in such cases an individual's privacy may not be respected. Thus within organizations, individuals' data should be restricted in terms of access and anonymous, by removing all information that can directly link data items to individuals via generalization or suppression before disclosing, so that privacy is not beached. Such a process is referred to as data anonymization.

A contemporary approach dealing with data privacy relies on k-anonymity. The k- anonymity model proposed by Samarati and Sweeney [18,23] is a simple and practical privacy-preserving approach to protect data from individual identification. The k-anonymity model works by ensuring that each record of a table is identical to at least $(k-1)$ other records with respect to a set of privacy-related features, called *quasi-identifiers*, that could be potentially used to identify individuals by linking these attributes to external data sets [14]. Therefore, privacy related information can not be revealed from the k-anonymity protected table during a data mining process. For example, consider the patient diagnosis records in a hospital in Table 1, where the attributes *ZipCode, Gender*, *Age* and *Education* are regarded as quasi-identifiers. A diagnosis classifier can predict the patient's illness history based on attributes of *ZipCode, Gender*, *Age* and *Education* using these data. If the hospital simply publishes the table to other organizations for classifier development, those organizations might extract patients' disease histories by joining this table with other tables [5]. By contrast, Table 2 is a 3-anonymization version where data values of Table 1 in attributes *ZipCode, Gender*, *Age* and *Education* have been generalized as common values and the number of records in its two equivalence classes are both equal to three. It should be noted that the value of k in the k-anonymity model is specified by users according to the purpose of their applications. By enforcing the k-anonymity requirement, it is guaranteed that even though an adversary knows that a k-anonymous table contains the record of a particular individual and also knows some of the quasi-identifier attribute values of the individual, he/she cannot determine which record in the table corresponds to the individual with a probability greater than $\frac{1}{k}$ [3]. This indicates that the larger the values of k, the less chance the adversary has of being able to determine personal identifiable information and the data is more protected. On the other hand, if the k-values are too large it incurs more information loss. Therefore, the k-value of the k-anonymization problem should not be too small or too large.

Usually, there are two methods to accomplish in k-anonymizing a dataset. The first one is suppression which involves not releasing an entire tuple or a value at all to the third party,

Table 1 Patients records in a hospital

ZipCode	Gender	Age	Education	Disease	Expense
4350	Male	24	9th	Flue	2,000
4351	Male	25	10th	Cancer	3,500
4352	Male	26	9th	HIV+	6,500
4350	Male	35	9th	Diabetes	2,000
4350	Female	40	10th	Diabetes	3,200
4350	Female	38	11th	Diabetes	2,800

Table 2 3-anonymization table

ZipCode	Gender	Age	Education	Disease	Expense
435⋆	Person	[21–30]	Educated	Flue	2,000
435⋆	Person	[21–30]	Educated	Cancer	3,500
435⋆	Person	[21–30]	Educated	HIV+	6,500
435⋆	Person	[31–40]	Educated	Diabetes	2,000
435⋆	Person	[31–40]	Educated	Diabetes	3,200
435⋆	Person	[31–40]	Educated	Diabetes	2,800

which is just like deleting them. The other one is generalization which involves replacing the value or tuple with a less specific but semantically consistent value. For example, suppose following five ages of individuals 51, 52, 53, 53, 55 exist. We can generalize attribute *Age* to age groups 50–55. On the other hand, we can also generalize them to an other set 5⋆. However, we can suppress the age values by ⋆. Intuitively, generalization is better than suppression because of extracting at least some information. Undoubtedly, anonymization is accompanied by information loss. In order to be useful in practice, the dataset should stay as informative as possible. Hence, it is necessary to consider deeply the tradeoff between privacy and information loss. To minimize the information loss due to k-anonymization, all records are partitioned into several groups such that each group contains at least k similar records with respect to the quasi-identifiers and then the records in each group are generalized or suppressed such that the values of each quasi-identifier are the same. Such similar groups are known as clusters. In the context of data mining, clustering is a useful technique that partitions records into clusters such that records within a cluster are similar to each other, while records in different clusters are most distinct from one another [14]. Thus, the k-anonymity model can be addressed from the viewpoint of clustering.

As discussed, a key difficulty of data anonymization comes from the fact that data quality and privacy are conflicting goals. Although it is possible to enhance data privacy by hiding more data values, it decreases data quality. By contrast, disclosing more data values increases data quality but decreases data privacy. Thus it is necessary to devise new k-anonymization approaches that best address both the quality and the privacy of the data. To overcome this challenge, this paper proposes a new clustering method for k-anonymization. This method has a time complexity of $O(\frac{n^2}{k})$ in the clustering stage, where n is the total number of records that contain individuals concerning their privacy. However, the algorithm requires sorting the tuples in the dataset once, which alone takes $O(n*logn)$ time. According to this method, first exclude the number of records containing individuals who do not bother about the disclosure of personal identification information. Sort all records by their quasi-identifiers and partition all records into $[\frac{n}{k}]$ groups. Randomly select a record r from the first group to form the first

cluster and the first records of the subsequent clusters will form in a systematic way. Then adjust the records in each group in a systematic way such that each group contains at least k records. Finally distribute the records of individuals who do not bother about the disclosure to their closest clusters or these records constitute another cluster/clusters depending on the number of such records and the k-value. Note that the process of including such records causes no information loss. There are many clustering based k-anonymization techniques in the literature [3,5,8,14,15]. However, the proposed systematic clustering method differs from previously proposed clustering based k-anonymization methods in four different ways. First, our method endeavours to make all clusters simultaneously. By contrast, the methods proposed by Byun et al. [3] and Loukides and Shao [15] build one cluster at a time. Second, it takes less time than the previous two methods as only the first record randomly selects and the subsequently records from in a systematic way. Third, since the first record of each cluster contains a non identical value, this method easily captures if there are any extreme values, and lastly the total information loss will be reduced as in the final step the process incurs no information loss. The performance of the proposed method is compared against the method proposed by Byun et al. [3]. The experimental results show that the proposed clustering method outperforms their method with respect to both information loss and computational efficiency.

The reminder of this paper is organized as follows. We present some concepts relating to information loss and a brief overview of the clustering based approaches for k-anonymization in Sect. 2. In Sect. 3 we present proposed systematic clustering method and the performance study is presented in Sect. 4. Usability of the proposed clustering algorithm in incremental datasets is illustrated in Sect. 5. Finally, concluding remarks are included in Sect. 6.

2 Preliminaries relating to k- anonymization

The k-anonymity model has drawn considerable interest in the research community for the last few years and a number of algorithms have been proposed [1,6,7,11,12,20–22]. However, these suffer from high information loss mainly due to reliance on pre-defined generalization hierarchies [1,7,11,22] or total order [6,12] imposed on each attribute domain. Some existing work on k-anonymization has attempted to capture usefulness by measuring the number of total suppressions [17], the size of the anonymized group [1,12], the height of generalisation hierarchies [3,18], or information loss through anonymization [25]. However, such metrics fail to detain security. In other works by Machanavajjhala et al. [16], and Truta and Vinay [16,24], attempts have been made to enhance protection by enforcing anonymized groups. The intuition behind this is that if the values of a sensitive attribute of an anonymized group are quite diverse, then it is difficult for an attacker to breach privacy. However, these frequency-based criteria treat numerical attributes as categorical and thus protection is not captured adequately. For instance, l-diversity proposed by Machanavajjhala et al. [16] requires a sensitive attribute to have at least l distinct values in an anonymized group. Please refer to Ciriani et al. [6] for a survey of various k anonymization approaches.

2.1 Information loss

Anonymization via generalization or suppression usually causes information loss. Now a natural question arises of how much information is lost due to anonymization. Thus the idea of information loss is used to measure the amount of information loss due to k-anonymization. There are various methods of conniving information loss [1,3,10,14,19]. The measurement

of information loss in this article is based on the description given by Byun et al. [3]. Please also refer to Byun et al. [3] for more details.

Let η denote a set of records with r numeric quasi-identifiers $N_1, N_2, \ldots, N_r$ and s categorical quasi-identifiers $C_1, C_2, \ldots, C_s$. Let $\Im = \{\Omega_1, \Omega_2, \ldots, \Omega_p\}$ be a partitioning of η, such that $\cup_{i=1}^{p} \Omega_i = \eta$, Ω_i and Ω_j $(i \neq j)$ are pair wise mutually exclusive. To generalize the values of each categorical attribute $C_i (i = 1, 2, \ldots, s)$, let τ_{C_i} be the taxonomy tree defined for the domain of C_i.

Consider a cluster Ω in η which consists of some numerical and categorical attributes. Let $N_{i_{max}}$, $N_{i_{min}}$ be the maximum and minimum values of the records in Ω and $\eta_{N_{i_{max}}}$, $\eta_{N_{i_{min}}}$ be the maximum and minimum values of the records in η with respect to numeric attribute $N_i (i = 1, 2, \ldots, r)$ and $\cup_{C_i}$ be the union set of values in Ω with respect to the categorical attribute $C_i (i = 1, 2, \ldots, s)$. Then the amount of information loss due to generalizing Ω, denoted by $IL(\Omega)$ is defined as

$$IL(\Omega) = \mid \Omega \mid . \left(\sum_{i=1}^{r} \frac{N_{i_{max}} - N_{i_{min}}}{\eta_{N_{i_{max}}} - \eta_{N_{i_{min}}}} + \sum_{j=1}^{s} \frac{H(\Lambda(\cup_{C_j}))}{H(\tau_{C_j})} \right)$$

where $\mid \Omega \mid$ is the number of records in Ω, $\tau(\cup_{C_j})$ is the subtree rooted at the lowest common ancestor of every value in $\cup_{C_j}$ and $H(\tau)$ is the height of taxonomy tree τ.

Suppose that the total number of records in η is partitioned into p clusters, namely $\Im = \{\Omega_1, \Omega_2, \ldots, \Omega_p\}$. The total information loss of η is the sum of the information loss of each $\Omega_i (i = 1, 2, \ldots, p)$. Therefore the total information loss will be:

$$\begin{aligned} IL(\eta) &= \sum_{i=1}^{p} IL(\Omega_i) \\ &= \sum_{i=1}^{p} \mid \Omega_i \mid . \left(\sum_{k=1}^{r} \frac{N_{ik_{max}} - N_{ik_{min}}}{\eta_{N_{ik_{max}}} - \eta_{N_{ik_{min}}}} + \sum_{j=1}^{s} \frac{H(\Lambda(\cup_{C_{ij}}))}{H(\tau_{C_{ij}})} \right) \end{aligned} \tag{2.1}$$

The main objective of clustering techniques is to construct the clusters in such a way that the total information loss of η will be at a minimum.

Example Consider patients records in Table 1 and the 3-anonymization table in Table 2. The anonymized table consists of two clusters. The first cluster consists of the first three records and the second clusters consists of the last three records. Consider attributes ZipCode, Gender, Age, Education, where Age is a quantitative variable and the others are categorical variables. Also consider the taxonomy tree of ZipCode, Education and Gender in Figs. 1, 2 and 3 respectively. In the table the number of clusters is 2 and the size of each cluster is 3. In the first cluster the maximum and minimum values are respectively 26 and 24 and in the second cluster these values are respectively 40 and 35. Also the maximum and minimum values of all records are respectively, 40 and 24. Then the total information Loss of the anonymized table in Table 2 will be

$$IL(\eta) = \mid 3 \mid \left(\frac{26 - 24}{40 - 24} + 1 + 1 + \frac{1}{2} \right) + \mid 3 \mid \left(\frac{40 - 35}{40 - 24} + 1 + 1 + \frac{2}{2} \right) \approx 14.81. \tag{2.2}$$

2.2 Clustering based techniques

Clustering based techniques are now used in k-anonymization to protect the privacy of sensitive attributes and there are various clustering techniques in the literature [3,5,12,14,15].

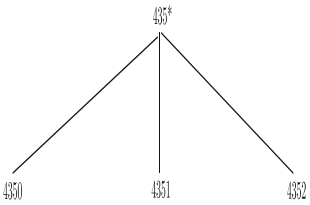

Fig. 1 Taxonomy tree of *ZipCode*

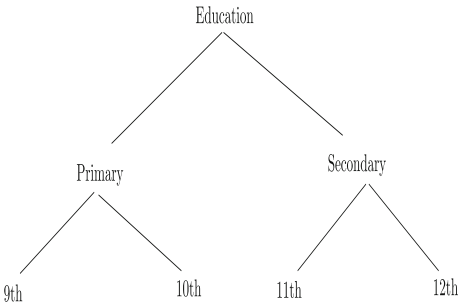

Fig. 2 Taxonomy tree of *Education*

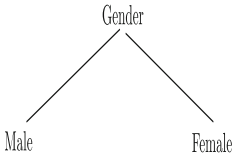

Fig. 3 Taxonomy tree of *Gender*

Byun et al. [3] introduced clustering techniques instead of equivalence class on k anonymization and proposed the greedy k-member clustering algorithm. This algorithm works by first randomly selecting a record r as the seed to start building a cluster, and subsequently selecting and adding more records to the cluster such that the added records incur the least information loss within the cluster. Once the number of records in this cluster reaches k, this algorithm selects a new record that is the furthest from r, and repeats the same process to build the next cluster. When there are fewer than k records not assigned to any cluster yet, this algorithm then individually assigns these records to their closest clusters. This algorithm has two drawbacks. First, it is slow. Second, it is sensitive to outliers. To build a new cluster, this algorithm chooses a new record that is the furthest from the first record selected for the previous cluster. If the data contains outliers, it is likely that outliers have a great chance of being selected. If a cluster contains outliers, the information loss of this cluster increases. The time complexity of the algorithm is $O(n^2)$, where n is the number of records in the data set to be anonymized. Their experimental results showed that the k-member algorithm causes significantly less information loss than another k-anonymization technique called "Mondrian" proposed by LeFevre et al. [12].

Loukides and Shao [15] proposed another clustering technique for k-anonymization. Similar to k-member this algorithm forms one cluster at a time. But, unlike the k-member algorithm, this algorithm chooses the seed of each cluster randomly. Also, when building a cluster, this algorithm keeps selecting and adding records to the cluster until the information loss exceeds a user defined threshold. If the number of records of a particular class is less than k, the entire cluster is deleted. With the help of the user-defined threshold, this algorithm is less sensitive to outliers. The time complexity of the algorithm is $O\left(\frac{n^2 log(n)}{c}\right)$, where c is the

average number of records in each cluster. However, this algorithm also has two drawbacks. First, it is difficult to decide a proper value for the user-defined threshold. Second, this algorithm might delete many records, which in turn cause a significant information loss. Chiu and Tsai [5] proposed another algorithm for k-anonymization that adapts the weighted feature c-means clustering. Unlike the previous two algorithms, this algorithm attempts to build all clusters simultaneously by first randomly selecting $\lfloor \frac{n}{k} \rfloor$ records as seeds. Then this algorithm allocates all records in the data set to their respective closest cluster and consequently updates feature weights to minimize information loss. This process is continued until the assignment of records to cluster stops changing. If some clusters contain fewer than k records, then those clusters should be merged with other large clusters to satisfy the k-anonymity requirement. One of the main drawback of this algorithm is that it can only be used for quantitative quasi-identifier. The time complexity of this algorithm is $O(\frac{t^2}{k})$, where t is the number of iterations needed for the assignment of records to clusters to converge.

To reduce the information loss and execution time recently Lin and Wei [14] proposed an efficient one-pass k-mean clustering problem that runs in $O(\frac{n^2}{k})$. They showed that their algorithm performs better than the proposed algorithm of Byun et al. [3] with respect to both execution time and information loss. Like Chiu and Tsai's [5] algorithm, this algorithm forms all clusters at a time. According to their methods first sort all records by their quasi-identifiers, then determine approximate number of clusters, by $p = \frac{n}{k}$, where k is the cluster size. Then randomly select p records as seeds to build p clusters. For each record r the algorithm finds the cluster that is closest to r, assigns r to that cluster and subsequently updates the center point. Finally, if some clusters contain more than k records remove excess records from those clusters that are dissimilar to most of the records and then add these records to other similar clusters (whose size is less than k). Although this method has less execution time there is still a chance of being affected by extreme values. Again if this algorithm first selects p records that come from the same equivalent class then the total information loss will be higher.

3 The new systematic clustering method

As discussed before, clustering escorts to better data quality of the disclosed dataset as it partitions a set of records into groups such that records in the same group are more similar to each other than to records of other groups. If the records in a particular group are more similar, the group leads to a minimal generalization and thus incurs less information loss. In this respect, the problem of k-anonymization can also be considered as a clustering problem, where each equivalent class is a cluster and the size of each cluster is at least k. So the optimal solution of a clustering problem is to construct a set of clusters such that the total information loss will be at a minimum. In this section, we formally define and present our systematic clustering algorithm that minimizes the information loss and respects the k-anonymity requirement.

3.1 Systematic clustering problem

There are various clustering problems in the literature. Among them, the k-center clustering problem proposed by Gonzalez [8] aims to find k clusters from a given dataset such that the maximum inter-cluster distance (or radius) is minimized. Thus the optimum solution is to constitute p clusters $\{\Omega_1, \Omega_2, \ldots, \Omega_p\}$ in such a way that it minimizes the cost metric

$$MAX_{i=1,\ldots,p} MAX_{j,k=1,\ldots,|\Omega_i|} D(r_{i,j}, r_{i,k}), \tag{3.3}$$

where $r_{i,j}$ represents a data point in cluster Ω_i and $D(x, y)$ is a distance between two data points, x and y.

In the k-anonymity problem the only restriction is that the number of records in each equivalence class should be at least k and there is no such restriction about the number of clusters. So a clustering problem is to form in such a way that each cluster contains at least k similar records and the sum of information losses of all clusters is minimum. The proposed k member clustering problem of Byun et al. [3] satisfies this criterion but one of the most important problems of this algorithm is that it spends a lot of time selecting records from the input set. To reduce time of selecting records from the whole set, a systematic method of selecting records may be helpful. To apply a systematic method of selecting records, first of all it is necessary to sort all records in the whole data set with respect to quasi-identifiers. For example, consider the dataset in Table 1, where there are six records and suppose that dataset is already sorted according to the quasi identifier attributes *ZipCode*, *Gender*, *Age* and *Education*. If the anonymized table follows 3-anonymity requirements, then the number of clusters should be $\frac{6}{3} = 2$. First select a record (say, the 2th record) from the first 3 records to form the first cluster. Then select $(2 + 3)$th $=$ 5th record in a systematic way to form the second cluster. Now again select another record from the first 3 records (say, 3rd, not 2th as it already selected) and calculate the information loss with both clusters using the Eq. (2.1). The information loss is 4.25 and 7.75, if this record is included in the first cluster and second cluster respectively. So, the 3rd record will be included in the first cluster as it causes the least information loss. Similarly, select $(3 + 3)$th $=$ 6th record in a systematic way and include it in the the second cluster. Finally select the 1st and $(1 + 3)$th $=$ 4th record and include these records respectively as first and second cluster as they will then cause least information loss. If the total number of records is not exactly divisible by the k-anonymity parameter, then the rest of the records will be included in the similar clusters where information loss is minimum and this process continues until the number of records in a particular cluster is k to satisfy the k-anonymity requirement. Thus we pretend the k-anonymity problem as a clustering problem, referred to as a systematic clustering problem.

Definition 1 (*Systematic clustering problem*) The systematic clustering problem is to find a set of clusters from a given set of n records such that each cluster contains at least k ($k \leq n$) records (where the records are selected in a systematic way and are included in a cluster that causes the least information loss) and that the sum of all intra-cluster distances is minimized. More specifically, if η be a set of n records and k the specified anonymization parameter, the optimal solution of the systematic clustering problem is a set of clusters $\Im = \{\Omega_1, \Omega_2, \ldots, \Omega_p\}$ such that:

1. $\Omega_i \cap \Omega_j = \Phi$, for all $i \neq j = 1, 2, \ldots, p$,
2. $\cup_{i=1,\ldots,p}^{p} = \eta$,
3. for all $\Omega_i \in \Im$, $\mid \Omega_i \mid \geq k$, and
4. the total information loss obtained by using Eq. (2.1) is minimized.

In Definition 1, a set of clusters are constructed in such a way that the clusters are mutually exclusive, the sum of records of all clusters is equal to the total number of records and the size of each cluster is at least k which satisfies the criteria of k-anonymization. The problem tries to minimize the sum of all intra-cluster distances, where an intra-cluster distance of a cluster is defined as the maximum distance between any two records in the cluster. In the following subsection we formally design a systematic clustering algorithm.

3.2 Systematic clustering algorithm

Based on the information loss in Eq. (2.1) and the definition of a systematic clustering problem, we are now ready to discuss a systematic clustering algorithm. The general idea of the algorithm is as follows.

Note that for collecting medical data from patients it may be expected that some patients are not concerned about the privacy of their medical records and the other attributes. We would like to explore this opportunity because unnecessary anonimization may produce more information loss. Let q be the probability that a particular patient is not concerned about the privacy of medical records. Then out of n patients we can expect that on average nq patients are not concerned about their privacy. According to this method first exclude the records of individuals who are not concerned about the privacy. Then sort all records by their quasi-identifiers and identify the equivalence class and the number of clusters by, $p = \frac{(n-nq)}{k}$, where k is the anonymity parameter for k-anonymization and round this as integer. Randomly select a record r_i from first k records as seeds to form the first cluster. If there are p clusters to be formed then select the $(r_i + k)$th, $(r_i + 2k)$th,…, $\{r_i + (p-1)k\}$th records in a systematic way to form the 2nd, 3rd,…, pth cluster respectively. Select another record $r_j (j \neq i)$ from the first k records and add this record to the cluster which causes least information loss. Similarly in a systematic way select $(r_j + k)$th, $(r_j + 2k)$th,…, $\{r_j + (p-1)k\}$th records and add these records to their respective clusters that cause least information loss. If any cluster size is exactly k, stop adding records to that cluster and continue the same process until all records of first k records are finished. If $(n - nq)$ is not exactly divisible by k and there are still some records left, add these records to their closest clusters that incur least information loss. Finally distribute the nq records to their closest clusters or these nq records constitute separate cluster/clusters depending on their size. Note that these nq records incur no information loss. Since only the first record randomly selects and the subsequent records from in a systematic way, it has less execution time. Again usually the first record of each cluster contains a non identical value, so this algorithm easily captures if there are any extreme values. Moreover, this algorithm is adding some records that contain no information loss, so it is a natural expectation that the total information loss will be reduced. The systematic clustering algorithm is shown in Table 3. In the algorithm it is assumed that all n individuals are concerned about their privacy.

Definition 2 (*Systematic clustering decision problem*) In a given data set of n records, there is a clustering scheme $\Im = \{\Omega_1, \Omega_2, \ldots, \Omega_p\}$ such that

1. $\mid \Omega_i \mid \geq k$, $1 < k \leq n$: the size of each cluster is greater than or equal to a positive integer k, and
2. $\sum_{i=1}^{p} IL(\Omega_i) < c, c > 0$: the total information loss of the clustering scheme is less than a positive integer c.

where each cluster $\Omega_i (i = 1, 2, \ldots, p)$ contains the records that are more similar to each other, such that they require minimum generalization and thus cause least information loss. In the following subsection we are going to discuss some properties of the proposed systematic clustering algorithm.

3.3 Properties of the proposed algorithm

As discussed before, the proposed algorithm is designed in such a way that it finds a solution of k-anonymization in a greedy manner. This algorithm stops adding records in a particular

Table 3 Systematic clustering algorithm

Input: a set η of n records containing individuals concerning their privacy, where $\eta_1, \eta_2, \ldots, \eta_n \in \eta$; the value k for k-anonymity
Output: a partitioning $\Im = \{\Omega_1, \Omega_2, \ldots, \Omega_p\}$ of τ
1. Sort all records in η by their quasi-identifiers;
2. Let p:= int$\lfloor \frac{n}{k} \rfloor$;
3. Get randomly k distinct records $r_1, r_2, \ldots, r_k$ from first 1 to k;
4. Let p_{ij} is the jth element in the ith cluster;
5. For $i = 1$ to p;
6. Let $p_{i1} := \eta_{[r_1+k(i-1)]}$;
7. Next i;
8. For $j := 2$ to k;
9. For $i := 1$ to p;
10. Let IL_i: = InfoLoss($\eta_{[r_j+k(i-1)]}$);
11. Let X:= Find cluster number with lowest IL_i;
12. where cluster size $\leq k$;
13. Add $\eta_{[r_j+p(i-1)]}$ to p_x;
14. Next i;
15. Next j;
16. Let $e := (n - pk)$;
17. Find extra element $E_1, E_2, \ldots, E_e \in E$;
18. For $k := 1$ to e;
19. For $m := 1$ to p;
20. Let $IL_m :=$ InfoLoss(E_k) in cluster m;
21. Next m;
22. Let $X :=$ Find cluster number with lowest IL;
23. Add E_k to p_x;
24. Next k;

cluster if the cluster size is exactly k. Again it always keeps in mind to add records that incur less information loss. Moreover, the records are selected in a systematic way that make the algorithm faster. With respect to this, this algorithm has the following desirable properties.

Theorem 1 *Let n be the total number of input records and k be the specified k anonymity parameter. The time complexity of the systematic clustering algorithm in the clustering stage is in* $O(\frac{n^2}{k})$

Proof After sorting the records with respect to the quasi-identifiers, the systematic clustering algorithm determine the numbers of clusters by $p = \frac{n}{k}$. Then it selects the records as seeds in a systematic way to form all p clusters simultaneously. Thus for each tuple in the dataset, the algorithm needs to assign it to one of the p clusters, which has a complexity of $O(p)$. As a result, the assignment of all tuples to the clusters has a time complexity of

$$\begin{aligned} T &= O(\text{Number of tuples} * \text{Number of clusters}) \\ &= O(n * p) = O\left(n * \frac{n}{k}\right) = O\left(\frac{n^2}{k}\right). \end{aligned} \tag{3.4}$$

Therefore, the total execution time is in $O(\frac{n^2}{k})$. □

Theorem 2 *Let n be the total number of input records and q be the probability that a particular individual doesn't bother about the disclosure. Then the systematic clustering algorithm in fact work out the information loss of* $(n - nq)$ *individuals instead of all n individuals.*

Proof If q be the probability that a particular individual does not bother about the disclosure, then out of n individuals, nq individuals are not bothered about the disclosure. Assume that these nq records are in one separate cluster that causes no information loss. Also let $IL(\eta)$ and $IL(\eta_{all})$ be the total information loss due to k-anonymization for a systematic clustering algorithm and any other clustering algorithm respectively. According to the systematic clustering algorithm, the total information loss will be:

$$\begin{aligned} IL(\eta) &= IL(n) \\ &= IL(nq) + IL(n - nq) \\ &= 0 + IL(n - nq) = IL(n - nq). \end{aligned} \tag{3.5}$$

Thus, the systematic clustering algorithm actually calculates the information loss of $(n - nq)$ records instead of calculating the information loss of all n records. □

Theorem 3 *Let n be the total number of input records and k be the anonymity parameter in k-anonymization. Then according to the systematic clustering algorithm, the cluster size of any cluster is at least k but no more than* $(2k - 1)$.

Proof Let n be the total number of input records. According to systematic clustering, first select the initial seeds of all clusters in a systematic way and subsequently select adding more records to the clusters such that the added records incur the least information loss. Again this algorithm stops adding records to a particular cluster if the number of records is exactly k. So in the worst case, if there are $(k - 1)$ records left and if all these records are included in a cluster that already contains k records, then the total number of records in that cluster will be $(k + k - 1) = (2k - 1)$. Therefore the maximum size of a cluster will be $(2k - 1)$. □

The properties discussed above show the utility of the proposed clustering algorithm with respect to both information loss and execution time. However it is necessary to check the efficiency of the algorithm by doing an experiment. In the following section the experimental results of the proposed algorithm are discussed.

4 Experimental results

The objective of our experiment is to investigate the recital of our approach in terms of data quality and computational efficiency. To accurately evaluate our approach, the performance of the proposed systematic clustering algorithm is compared in this section with the k-member algorithm [3]. Byun et al. [3] showed that a k-member algorithm causes significantly less information loss than Mondrian, proposed by LeFevre et al. [12]. As it already evaluated that the k-member algorithm outperforms Mondrian, in this paper we compare our proposed algorithm with the k-member algorithm. Both experiments are implemented with Excel VBA programming language and run on a 3.20 GHz Pentium (R) D CPU processor machine with 2 GB of RAM. The operating system on the machine was Microsoft Windows XP Professional Version 2002.

We utilized Adult dataset from the UC Irvine Machine Learning Repository [9] for both the experiments. It should be noted that the Adult dataset is considered as a standard benchmark for evaluating the performance of any k-anonymity algorithm. We deleted the records

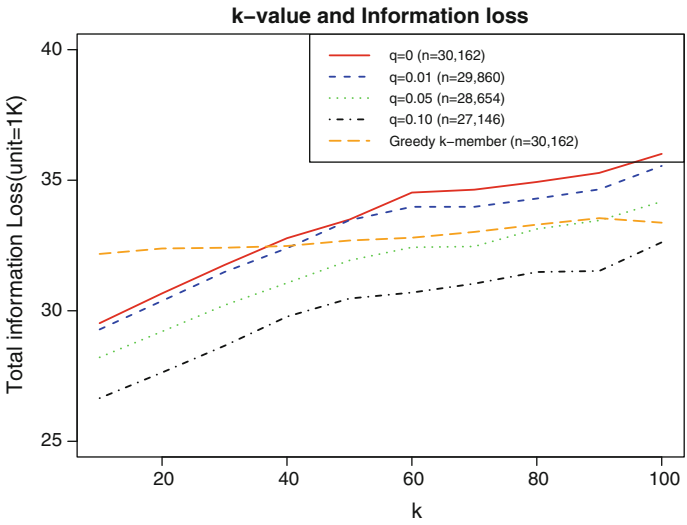

Fig. 4 Information loss

with missing values and retained only three of the original attributes, namely Age, Sex and $Education$ as quasi-identifiers. Among these, Age is a numeric attribute but Sex and $Education$ are the categorical attributes. The taxonomy trees for these two categorical attributes are based on those defined in [7].

The experiments are conducted as follows. First, the systematic clustering algorithm with three different scenarios ($q = 1\%$, $q = 5\%$, $q = 10\%$) and the k-member algorithm are run five times for every k value, and total information loss and execution time are collected for each run. Then, the average of every five runs using the same algorithm and k is computed and reported here.

Figure 4 shows the information loss of both the systematic clustering algorithm along with three levels ($q = 1\%$, $q = 5\%$, $q = 10\%$) and the k-member algorithm [3] with respect to the k-anonymity parameter. It shows that the total information loss of each of the algorithms increases as k increases. The logic behind this is that as k increases the clusters need to maximum generalization and this incurs more information loss. One of the most important criteria of choosing a best clustering method is that it causes the least information loss. In this aspect, a systematic clustering method with $q = 10\%$ uniformly satisfies this criterion. That means that if at least 10% of individuals do not care about the disclosure then the systematic clustering method is the best choice as a clustering technique for k-anonymization in the data mining environment. However, as Fig. 4 shows for some moderate values of k ($k \leq 40$), the systematic clustering method always incurs less information loss even if all individuals are concerned about their privacy. In practice, the k-value of the k-anonymization problem should not be too small or too large as small values of k signify higher probability of disclosure and large values of k signify the more information loss. Thus in a realistic situation, $k \leq 40$ is reasonable for k-anonymity problems and in that case the proposed systematic clustering algorithm attains a reasonable dominance over the k-member algorithm.

On the other hand, Fig. 5 displays the execution time of both algorithms. Figure 5 clearly shows that the execution time of the proposed systematic clustering algorithm with all different scenarios is much less than in the k-member algorithm. The greedy k- member algorithm takes too much time as it spends a lot of time selecting records from the input set. Again

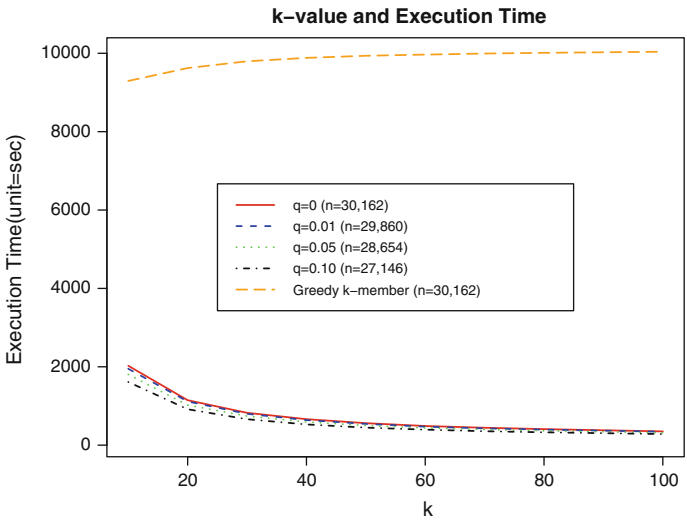

Fig. 5 Execution time

as expected, the execution time of the systematic clustering algorithm decreases with the increase of the probability that a particular individual does not care about the privacy as in this case the total number of records in the input set decreases. Thus it can be said that the proposed method is superior to the k-member algorithm in terms of both information loss and execution time.

As discussed before, a main challenge in data mining is to enable the legitimate usage and sharing of mined information while at the same time guaranteeing proper protection of the original sensitive data. Because of increasing concerns about the privacy of individuals, privacy preserving is an important issue and has captured the attention of many researchers in the data mining research community. Although k-anonymity is a proper solution of protecting sensitive attributes in a dataset, the main drawback of the method is the information loss. Thus, a natural question arises in this case: how to design a method in such a way that causes less information loss and execution time and at the same time satisfies the k-anonymity requirement. Based on this, an algorithm is developed in this paper that uses the idea of clustering and incurs as little information loss as possible. As Figs. 4 and 5 show, the proposed algorithm causes less information loss and execution time, and it demonstrates the flexibility and the usability of the proposed algorithm.

5 Anonymization for incremental datasets

Anonymization based on k-anonymity models has been the focus of intensive research in the last few years. However the current techniques related to the k-anonymity model are limited only where it is assumed that the entire dataset is available at the time of release (static data). This assumption leads to a shortcoming as data nowadays are continuously collected (thus continuously growing) and there is a strong demand for up-to-date data at all times [4]. In such a dynamic environment the proposed systematic clustering method can be easily applied without any modification of the previous anonymously released data.

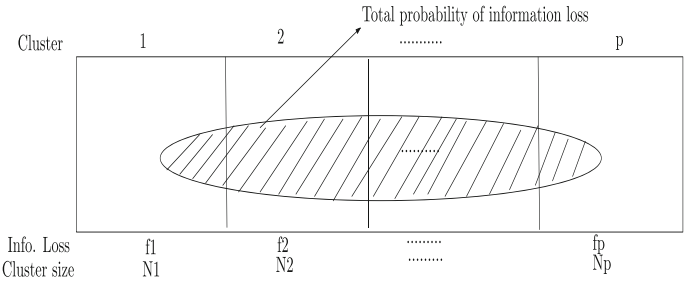

Fig. 6 Bays approach

Suppose that a hospital wants to publish its patients for medical researchers and it is assumed that the hospital has already released the entire static data by using the systematic clustering method. The hospital can then infer the released data and this prior information can be used when a new record will be released. Assume that the hospital anonymized n records containing individuals concerning their privacy and build p clusters, where the information loss and the size of the ith cluster are respectively as f_i and N_i $\left(\sum_i^p N_i = n\right)$. Thus, the total information loss is $f = \sum_i^p f_i$ and the proportion of information loss in the ith cluster is $P(f|i) = \frac{f_i}{n}$. Moreover, the probability that a new record will be included in the ith cluster is $P(i) = \frac{N_i}{n}$. Thus according to the Bays approach, the total probability of information loss is

$$P(f) = P(1)P(f|1) + P(2)P(f|2) + \cdots + P(p)P(f|p) = \sum_{i=1}^{p} \frac{N_i f_i}{nf} \tag{5.6}$$

The Bays approach is illustrated in Fig. 6. Now suppose that the hospital wants to release a new record and assume that this record contains individuals' concerning their privacy (so needs anonymization and thus produce information loss), then the probability that this record will be included in the ith cluster is

$$P(i|f) = \frac{\frac{N_i f_i}{nf}}{\sum_{i=1}^{p} \frac{N_i f_i}{nf}} = \frac{N_i f_i}{\sum_{i=1}^{p} N_i f_i} \tag{5.7}$$

The higher probability indicates that the information loss will be higher if the new record is included in that particular cluster. So the new record should be included where the posterior probability is at a minimum. However, this is a preliminary idea of including a new record in a cluster. The easiest way to calculate the information loss of the new record with the existing clusters is to include the record with the cluster that causes the least information loss. As the preconstructed clusters based on the systematic clustering algorithm satisfy the k-anonymity requirement, the inclusion of the new record also respects the condition without any modification of preexisting clusters. Thus without any loss of generality the systematic clustering algorithm can be used for incremental datasets.

6 Conclusion and future works

In this paper, we have proposed an efficient algorithm for k-anonymization to minimize the information loss during the anonymization process and assure data quality. The proposed

technique uses the idea of clustering and we refer to this as the systematic clustering algorithm. The basic concepts of the proposed algorithm were discussed and investigated through example and properties. The time complexity of the developed algorithm is in $O(\frac{n^2}{k})$, where n is the total number of records containing individuals concerning their privacy. Finally a comparison was made on the proposed algorithm with the k-member algorithm proposed by Byun et al. [3] through experiment. For any k-anonymization algorithm, there are two significant criteria to judge the superiority of the algorithm, namely, information loss and execution time. The experimental results show that the proposed systematic clustering algorithm has a reasonable dominance over the k-member algorithm. This shows the utility and the efficiency of the proposed clustering algorithm. Finally a way out was shown to be used for continuously growing data without any violation of the k-anonymity requirement.

Recently there are many disparities of the k-anonymity model have been proposed in the literature to further protect the private information, e.g., l-diversity [16], t-closeness [13], (α, k)-anonymity [26]. Our further work is to extend the systematic clustering algorithm to these models.

References

1. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: International Conference on Data Engineering (2005)
2. Byun, J.W., Bertino, E.: Micro-views, or on how to protect privacy while enhancing data usability: concepts and challenges. SIGMOD **35**(1), 9–13 (2006)
3. Byun, J.W., Kamra, A., Bertino, E., Li, N.: Efficient k-anonymization using clustering techniques. In: International Conference on Database Systems for Advanced Applications (DASFAA) (2007)
4. Byun, J.W., Sohn, Y., Bertino, E., Li, N.: Secure anonymization for incremental datasets. In: 3rd VLDB Workshop on Secure Data Management (SDM) (2006)
5. Chiu, C.-C., Tsai, C.-Y.: A k-anonymity clustering method for effective data privacy preservation. In: Third International Conference on Advanced Data Mining and Applications (ADMA) (2007)
6. Ciriani, V., di Vimercati, S.D.C., Foresti, S., Samarati, P.: k-anonymous data mining: a aurvey. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining: Models and Algorithms, pp. 103–134. Kluwer Academic Publishers, Boston (2008)
7. Fung, B.C.M., Wang, K., Yu, P.S.: Top-down specialization for information and privacy preservation. In: International Conference on Data Engineering (2005)
8. Gonzalez, T.Z.: Clustering to minimize the maximum intercluster distance. Theor Comput Sci **38**, 293–306 (1985)
9. Hettich, C.B.S., Merz, C.: UCI repository of machine learning databases (1998)
10. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: SIGKDD (2002)
11. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Incogniti: efficient full-domain k-anonymity. In: ACM International Conference on Management of Data (2005)
12. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: International Conference on Data Engineering (2006)
13. Li, N., Li, T.: t-closeness: privacy beyond k-anonymity and l-diversity. In: ICDE (2007)
14. Lin, J.L., Wei, M.C.: An efficient clustering method for k-anonymization. In: Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society (2008)
15. Loukides, G., Shao, J.: Capturing data usefulness and privacy protection in k-anonymisation. In: Proceedings of the 2007 ACM Symposium on Applied Computing (2007)
16. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramanian, M.: l-diversity: privacy beyond k-anonymity. In: ICDE (2006)
17. Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: PODS, pp. 223–228 (2004)
18. Samarati, P.: Protecting respondent's privacy in microdata release. TKDE, **13**(6) (2001)
19. Solanas, A., Sebe, F., Domingo-Ferrer, J.: Micro-aggregation-based heuristics for p-sensitive k-anonymity: One step beyond. In: International Work-shop on Privacy and Anonymity in the Information Society (2008)
20. Sun, X., Li, M., Wang, H., Plank, A.: An efficient hash-based algorithm for minimal k-anonymity. In: ACSC, pp. 101–107, (2008)

21. Sun, X., Wang, H., Li, J.: Priority driven K-Anonymisation for privacy protection. In: AusDM, pp. 73–78 (2008)
22. Sweeney, L.: Achieving *k*-anonymity privacy protection using generalization and supression. Int. J. Uncertainty Fuzziness Knowledge-Based Syst. **10**(5), 571–588 (2002)
23. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowledge-Based Syst. **10**(5), 557–570 (2002)
24. Truta, T., Vinay, B.: Privacy protection: *p*-sensitive *k*-anonymity property. In: International Workshop on Privacy Data Management (PDM), p. 94 (2006)
25. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.C.: Utility-based anonymization using local recording. In: KDD 2006, pp. 785–790 (2006)
26. Wong, R.C.-W., Li, J., Fu, A.W.-C., Wang, K.: (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2006)